Gestão Semântica de Dados Meteorológicos Apoiados Por Ontologias de Proveniência

Sérgio Manuel Serra da Cruz[1,3,4] Tiago Marcos Barbosa[1], Ednaldo Oliveira dos Santos[2], Gustavo Bastos Lyra[2]

[1] Instituto de Ciências Exatas – Universidade Federal Rural do Rio de Janeiro (UFRRJ)
[2] Instituto de Florestas – Universidade Federal Rural do Rio de Janeiro (UFRRJ)
[3] Pós-Graduação em Modelagem Matemática e Computacional (PPGMMC/UFRJ)
[4] Programa de Educação Tutorial – (PET-SI/UFRRJ)

{serra,edmeteoro,gblyra}@ufrrj.br

Resumo. *As atividades na área de Meteorologia envolvem o processamento e gestão de grandes volumes de dados semiestruturados coletados a partir de diversas fontes. Assim, é fundamental que o profissional da área utilize ferramentas computacionais para auxiliá-lo na gestão destes dados. Este trabalho apresenta uma abordagem para organização e avaliação de dados meteorológicos baseado em conceitos de Proveniência e Web Semântica. Nele se discute a construção de uma ontologia desenvolvida a partir de ontologia de fundamentação UFO para apoiar experimentos meteorológicos que manipulam grandes volumes de dados.*

Abstract. *Meteorology handle large volumes of semi-structured data collected from various sources. Thus, it is crucial that the meteorologists use computational tools to assist in the management of these datasets. This work presents an approach to aid them to organize and evaluate meteorological data based on concepts of Semantic Web and data provenance. The article present a semantic tool and a well-founded ontology developed from UFO foundational ontology to support experiments that manipulate large amounts of curated data.*

1. Introdução

Atualmente, os centros operacionais e de pesquisa em prognósticos numéricos e diagnósticos de tempo e clima trabalham com grandes quantidades de dados complexos, semiestruturados e multivariados, tendo que interpretá-los num curto espaço de tempo. Portanto, um dos principais desafios dos estudos nas áreas de meteorologia e climatologia é gerar dados meteorológicos de qualidade.

Tanto a falta de dados confiáveis, quanto o excesso, são problemas de difícil tratamento. Os problemas das inconsistências e variações nas séries de dados podem ocorrer devido a inúmeros motivos, como por exemplo, erro humano na coleta, condições do instrumento de medida (manutenção e calibração), processamento errôneo dos dados, variações no intervalo de observações e às mudanças no ambiente circundante. Em Meteorologia, não é suficiente apenas medir, é necessário processar,

corrigir e garantir consistência temporal-espacial aos dados medidos da maneira mais eficiente possível, pois estes dados devem ser mantidos para as futuras gerações.

Os dados brutos podem ser armazenados em granularidades e estruturas físicas distintas (textos, planilhas, bancos relacionais, XML, LOD e RDF) e expressam conceitos e semânticas que podem variar no tempo. Neste contexto, o uso de ferramentas baseadas em ontologias que apoiem a gestão adequada dos dados e metadados dos experimentos em Meteorologia se faz cada vez mais necessário. A gestão dos dados e de proveniência tem como objetivo garantir alinhamento semântico dos conceitos utilizados por grupos geograficamente e temporalmente distintos, mas também melhorar a consistência, a integridade, organização, reprodutibilidade e a confiabilidade dos prognósticos.

Este trabalho tem como objetivo apresentar uma versão ampliada da ontologia Meteoro, apoiada no metamodelo PROV-DM da W3C [3] e nas ontologias de proveniência OvO (*Open proVenance Ontology*) [2] e *Semantic Sensor Network* (SSN) [13]. O trabalho se alinha com a temática da proveniência em *e-Science*, adota como referencial teórico a metamodelagem baseada na ontologia de fundamentação UFO (*Unified Foundational Ontology*) concebida por [4]. Como prova de conceito modelou-se uma ontologia em OntoUML e seu artefato computacional em OWL e um banco de dados que permitem aos meteorologistas desenvolverem consultas em linguagem SPARQL para avaliar a qualidade dos dados dos experimentos meteorológicos.

Este trabalho está organizado da seguinte forma: Seção 2 contém a fundamentação teórica e trabalhos relacionados. Seção 3 apresenta a abordagem proposta. Seção 4 apresenta detalhes da proposta e seus experimentos, e por fim, a Seção 5 conclui o artigo e apresenta limitações e perspectivas de trabalhos futuros.

2. Fundamentação Teórica

2.1.Dados Meteorológicos e Pré-Processadores

Historicamente identificam-se três eras na coleta de dados na Meteorologia. Na era "sinótica", os dados esparsos relacionados às medidas de superfície, coletados por instrumentos simples e transmitidos via correio ou telégrafo. Na era "radiossonda" as medidas no espaço tridimensional (superfície e altitude) eram realizadas por sensores acoplados em balões meteorológicos. Os conjuntos de dados cresceram, mas ainda eram esparsos. Atualmente, vive-se a era do "big data" [5], onde dados são coletados intensivamente por diversos sensores de alta resolução, câmeras, satélites, radares, boias, balões, embarcações, aviões e estações meteorológicas em alta frequência [8][6].

Os dados brutos consistem em medidas coletadas ou calculadas correspondentes aos elementos ou fenômenos atmosféricos, tais como precipitação, temperatura do ar e do solo, pressão atmosférica, entre outros. Eles são de natureza semiestruturada e estão associados a localizações espaço-temporais (tempo cronológico, latitude, longitude e altitude). Medidas são armazenadas em bancos de dados brutos, sendo posteriormente assimiladas por sistemas de pré-processadores [1][7] para em seguida serem analisadas. O sistema de pré-processadores utilizado neste trabalho foi desenvolvido em nosso grupo de pesquisas por Lemos Filho et al. [7].

2.2. Proveniência de Dados

A gestão da proveniência em experimentos científicos tem por objetivo auxiliar na busca de respostas as inúmeras indagações feitas pelos pesquisadores. Portanto, para que os dados meteorológicos sejam computados adequadamente, compartilhados e reutilizados com sucesso, é preciso assegurar que não só sejam livres de falhas e confiáveis, mas também anotados com metadados de proveniência retrospectiva, que captura as tarefas executadas sobre os dados e parâmetros utilizados, além das informações sobre o ambiente utilizado para derivar um resultado [9].

2.3 Ontologias de Fundamentação e de Proveniência

As ontologias de fundamentação descrevem conceitos gerais tais como: tempo, espaço, matéria, objeto, evento e ações e são independentes de domínio. Dentre as principais se destacam a SUMO, DOLCE e UFO [4]. Neste trabalho utilizou-se a técnica de metamodelagem de ontologias de domínio, tendo como referencial a ontologia UFO [8] para a publicação de dados, suporte de explicitação de compromissos ontológico dos modelos conceituais subjacentes e interoperabilidade semântica. Adotou-se apenas o fragmento UFO-A por que ela sistematiza os conceitos de tipos e estruturas taxonômicas, relações do tipo todo-parte, espaços de valores de atributos e propriedades relacionais. A Tabela 1 destaca as principais categorias formais da UFO-A e representa a inferência das categorias de tipo a partir das quatro metapropriedades da ontologia utilizada neste trabalho.

Tabela 1. Categorias de tipos de objetos e metapropriedades da UFO-A, adaptado de [8].

Tipo	Princípio Identidade	Identidade	Rigidez	Dependência
SORTAL	-	+		
<<Kind>>	+	+	+	-
<<subkind>>	-	+	+	+
<<role>>	-	+	-	+
<<phase>>	-	+	-	-
NON-SORTAL	-	-		
<<category>>	-	-	+	-
<<roleMixin>>	-	-	-	+
<<mixim>>	-	-	~	-

O metamodelo PROV-DM da W3C [3] se estabeleceu como uma especificação que possibilita a interoperabilidade da proveniência retrospectiva na Web. PROV-DM possui estruturas que permitem a modelagem dos conceitos essenciais para a representação da proveniência retrospectiva através de três tipos básicos (entidade, agente e atividade). Neste trabalho, limitou-se o escopo da captura e representação deste tipo de proveniência, a ser coletada na execução dos pré-processadores e sobre as transformações e correções de dados realizadas pelos meteorologistas.

Os modelos de proveniência mais recentes utilizam conceitos da área de Web Semântica, tanto para representar, quanto para consultar informações de proveniência. Por exemplo, em Moreau et al. [11] é definida uma ontologia denominada *Open Provenance Model Ontology* (OPMO) para capturar os conceitos baseados no metamodelo OPM (versão 1.1) e as inferências válidas. OPMO especifica a serialização RDF da versão 1.1 do modelo abstrato OPM. Recentemente, a ontologia PROV-O foi homologada pela W3C, e se caracteriza por descrever o conjunto de classes, propriedades e restrições do metamodelo PROV-DM. PROV-O é uma ontologia leve descrita em OWL2, que pode ser adotada em ampla gama de aplicações, sendo capaz de

representar informações de proveniência na Web. No entanto, essas ontologias de proveniência não estão alinhadas com as estruturas taxonômicas oferecidas por ontologias de fundamentação. A ontologia OvO [2] é uma ontologia bem fundamentada, que possui como suporte teórico a UFO para os conceitos relacionados com os descritores de proveniência dos experimentos científicos. OvO é uma ontologia de domínio que tem como pilares o metamodelo de proveniência PROV e a experimentação científica do tipo *in silico*.

2.4. Trabalhos Relacionados

Atemezing et al. [12] propõe uma estratégia para transformar dados meteorológicos em dados abertos, usando para isso uma ontologia própria. Eles enfatizam a importância e dificuldades de gerenciar grandes volumes de dados gerados por experimentos meteorológicos. As diferenças entre a nossa proposta e este trabalho são flagrantes. Primeiramente, diferente da ontologia Meteoro, os autores não usam uma ontologia de proveniência para tratar informações históricas dos processos e dados manipulados pelo pesquisador. Isso representa um indício de que a ontologia Meteoro poderá ser utilizada em ferramentas semânticas capazes de avalizar a consistência, integridade, reprodutibilidade e a confiabilidade dos experimentos. Além disso, Meteoro é apoiada em ontologias de fundamentação, o que confere compromissos ontológicos explícitos através de bases teóricas sólidas, melhorando a qualidade, oferecendo suporte lógico à modelagem do domínio.

Além das ontologias supracitadas, incorpora-se conceitos da SSN [13], que descreve sensores, observações ligado as à coleta de dados ambientais. No entanto, ela não descreve os conceitos de um domínio específico, por exemplo, não incorpora os conceitos de tempo, nem de localização das estações. O uso dessa ontologia como parte integrante da Meteoro é vantajoso, pois faz reaproveitamento dos conceitos e permite que *datasets* sejam interligadas mais facilmente através da propriedade *sameAs*.

3. Abordagem Proposta

Essa seção apresenta a ontologia de aplicação denominada Meteoro.

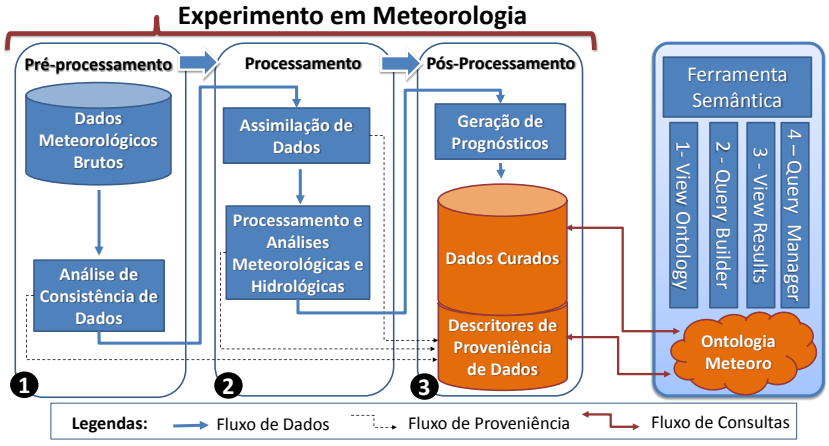

Figura 1. Representação conceitual de um experimento em Meteorologia e das camadas de uma ferramenta semântica.

A Figura 1 ilustra as etapas de geração de dados curados e anotados com proveniência em experimentos de meteorologias que manipulam grandes volumes de dados. Por limitação de escopo apenas os elementos gráficos em destaque (cor ocre) serão discutidos em detalhes neste artigo.

3.1 Experimentos em Meteorologia

Em geral, os prognósticos (experimentos) de precipitação pluvial na área da Meteorologia requerem longas séries de dados curados. Conceitualmente, eles são divididos em várias etapas, sendo que as três primeiras são: pré-processamento dos dados brutos, processamento e pós-processamento [1][7]. Esses experimentos por sua vez podem produzir novos conjuntos de dados que poderão ser explorados por novos conjuntos de ferramentas e pesquisadores.

3.2 Ferramenta Semântica

Uma ferramenta semântica no domínio das Meteorologias permite aos pesquisadores manipularem os conceitos associados aos dados e seus descritores de proveniência. Ela é composta pelas camadas:

1- ***Ontology Viewer*** - conversão do modelo ontológico (representação conceitual da ontologia) para linguagem OWL, representação dos mapeamentos dos esquemas das fontes de dados e de proveniência; A camada permite a visualização dos conceitos da ontologia como uma hierarquia de conceitos. 2- ***Query Builder*** - formulação visual de consultas em linguagem SPARQL que são enviados para o módulo processador de consultas (*Query Manager*). A formulação considera a necessidade de vários tipos de usuários e pode ser feito de duas maneiras: (1) para usuários com pouca experiência na linguagem SPARQL com base em visualização direta dos conceitos da ontologia mapeados a partir do banco de dados; (2) para usuários com experiência em SPARQL, é possível codificar manualmente as consultas na interface. 3- ***Results Viewer*** – organização e recarrega das consultas SPARQL. A camada também os resultados das consultas. 4- ***Query Manager*** - processamento das consultas formuladas pelo pesquisador. Além disso, realiza a integração entre os conceitos das ontologias com os repositórios de dados e de proveniência.

3.3 Ontologia Meteoro

Meteoro é uma ontologia de aplicação que tem sua modelagem conceitual apoiada a partir das ontologias de fundamentação UFO-A e da ontologia de proveniência OvO. Os modelos conceituais da Meteoro são representados através da linguagem OntoUML [8]. Esta característica facilita a modelagem e representa um formalismo simples de ser compreendido por profissionais que não sejam da área da computação. Na modelagem da Meteoro utilizou-se apenas o fragmento UFO-A, que trata especificamente de *endurants* (elementos que perduram no tempo) e conceitos importantes e distintos: indivíduos (Particular) e tipos (Universal).

Indivíduos são entidades que existem na realidade e possuem um conjunto específico de identidades únicas. Tipos, são padrões de características que podem ser instanciados em um número de diferentes indivíduos. Outro conceito importante para a compreensão do trabalho é a adoção de sortais. Sortais carregam princípios de identidade e individuação e contagem. Sortais podem ser do tipo rígido ou antirrígido [4].

A Figura 2A ilustra as principais classes da ontologia Meteoro. Os conceitos relacionados à subontologia de proveniência retrospectiva da OvO estão representados na verde, em amarelo os conceitos da subontologia de experimentos científicos em larga escala, também da OvO. Na cor azul estão os conceitos (variáveis meteorológicas) que serão mapeados pela ontologia Meteoro ou reaproveitados da ontologia de domínio SSN. Por exemplo, os dados são artefatos meteorológicos (*Artificat*) e seus subtipos podem ser os parâmetros de entrada e saída. Especificamente, pode-se dizer que os dados meteorológicos se dividem basicamente em variáveis medidas pelos sensores das estações meteorológicas. Por exemplo, valores de umidade do ar (*Humidity_Data*), precipitação (*Precipitation_Data*), vento (*Surface_Wind_Data*), radiação solar (*Solar_Radiation_Data*), pressão atmosférica (*Atmospheric_Pressure_Data*) e temperatura (*Temperature_Data*), este último se dividindo em duas partes: temperatura do solo (*Soil_Temperature*) e temperatura do ar (*Air_Temperature*). Esses conceitos são modelados pelo estereótipo <*SubKind*>, portanto possuem identidades próprias e herdam as propriedades da classe *Artifact*. Nesse caso, os subtipos de dados de entrada e saída foram mapeados como um sortal do tipo <*SubKind*>, justamente por serem únicos, mas ao mesmo tempo herdarem características em comum. Por exemplo, todo dado de entrada possui um valor de medição, ou uma data de medição, além disso, segundo nosso modelo são gerados por uma única estação meteorológica, o que também justifica o reaproveitamento do relacionamento *isGeneratedBy* que é mapeada pela relação formal criada entre a estação e os dados meteorológicos.

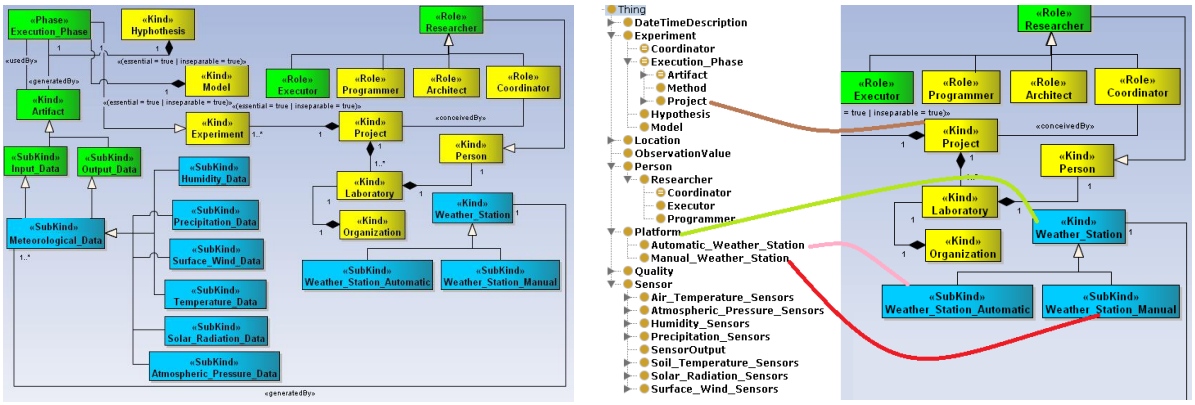

Figura 2A. Fragmento do Esquema da Ontologia Meteoro Modelado na Ferramenta *Enterprise Architect* usando o plug-in OLED da OntoUML. Figura 2B. Exemplificação do mapeamento dos conceitos da ontologia para OWL.

As classes representadas na cor verde ou amarela indicam descritores de proveniência (retrospectiva e prospectiva) relacionados aos experimentos em meteorologia. Os conceitos *Laboratory* e *Organization* são sortais rígidos e relacionam-se por agregação e composição com os conceitos subjacentes. O conceito *Laboratory* identifica unicamente os pontos no espaço geográfico, onde um projeto de pesquisa é concebido e onde seus experimentos científicos são conduzidos; é importante ressaltar que um projeto de pesquisa está associado com laboratório, sendo possível diferenciar em quais pontos os experimentos de um projeto são conduzidos. Por esses motivos, o conceito *Laboratory* requer identificação única, o que demanda o uso do tipo <*Kind*>.

O conceito *Laboratory* apresenta uma segunda característica importante, explicita a sua relação por agregação com o conceito Person que também é representado por um sortal

rígido <*Kind*>, e que identifica unicamente um pesquisador (*Researcher*). Aqui assume-se que ser pesquisador é uma propriedade extrínseca de uma pessoa, i.e., existem mundos em que uma pessoa não é pesquisador, todavia, ele permanece sendo uma pessoa. Da mesma forma, uma pessoa pode deixar de fazer parte de um laboratório de pesquisa ou mesmo de uma organização sem, no entanto, deixar de ser uma pessoa. O conceito *Person* representa todo o qualquer ser humano, no entanto, representou-se aqui apenas aqueles que desempenham algum papel (*role*) na condução de um experimento científico em meteorologia.

Para representar os possíveis papéis desempenhados por pessoas cujo atributo principal é ser pesquisador, adotou-se o tipo <*Role*> da UFO para representá-los como sortais antirrígidos. Para descrever esta sucessão de estereótipos de modo ontologicamente correta, adotou-se o *design pattern* apresentado por [8]. No caso específico, mapeou-se o conceito *Researcher* como sortal do tipo <*Role*>, pois em suas extensões ele possui indivíduos (pessoas) que pertencem a diferentes grupamentos humanos e que possuem seus princípios de identidade. Neste caso, um pesquisador é um não sortal. Cada instância de pesquisador deve ser uma instância de pessoa que carrega seu princípio da identidade. Por exemplo, neste caso definiram-se quatro sortais: executor, programador, arquiteto e coordenador (*Executor*, *Programmer*, *Architect* e *Coordinator*) como subtipo de *Researcher*. Por seu turno, estes sortais possuem princípios de identificação que são fornecidos pela classe *Person*. Resumindo, se α é um pesquisador (classe abstrata) então α deve ser uma instância de exatamente um subtipo de pesquisador, que possui o princípio da identidade fornecido por um sortal apropriado.

A codificação do artefato ontológico em linguagem OWL é uma etapa necessária para que as consultas SPARQL sejam processadas pelo computador. Na Figura 2B verificam-se exemplos onde os sortais do tipo <*Kind*>, representando a classe *Weather_Station*, que representa as estações meteorológicas e seus subtipos <*subkind*> e a esquerda a classe OWL *Plaftform* que foram mapeados da ontologia SSN, além de suas respectivas subclasses (*Manual_Weather_Station* e *Automatic_Weather_Station*). Observa-se ainda, o mapeamento da classe *Project*, da ontologia OvO. Note que essa transposição deveria ocorrer automaticamente, porém até o momento da execução do trabalho a linguagem OntoUML ainda não havia apresentava esse recurso de forma satisfatória para conversão em OWL.

3.5 Bancos de Dados e de Proveniência

A carga das séries de dados brutos foi operacionalizada pelos meteorologistas através da execução dos *workflows* de pré-processadores (detalhes em [7, 8]). Durante as execuções se armazenam os dados brutos e curados, assim como os descritores de proveniência retrospectiva de baixa granulosidade (relativos às execuções dos *workflows*), isto é, para cada arquivo de dados brutos carregado, são armazenados no repositório relacional os dados originais, os dados transformados (curados), os registros com os metadados de proveniência e os agentes (usuários) envolvidos no processo, além das informações de *timestamp* sobre as transformações. Diferentemente de nosso trabalho anterior [8], a versão atual do modelo de dados é compatível com a especificação PROV-DM, armazena metadados sobre atividades (pré-processadores e transformações de dados); entidades (dados brutos, curados, estações, laboratório,

experimento e projeto de pesquisa) e agentes (usuários e pesquisadores envolvidos no projeto e na operação do sistema).

As consultas SPARQL sobre os dados e proveniência permitem que meteorologias avaliem o processo de transformação dos dados. Além disso, atuam como um componente adicional para assegurar a qualidade dos dados currados. Eles mantêm informações sobre os dados utilizados e processos que o derivaram, agentes envolvidos na derivação. A Figura 3 ilustra apenas as principais classes do modelo conceitual do banco de dados necessárias à compreensão deste trabalho. Observe que no texto (ao lado do nome das classes) existe uma indicação gráfica de correlação da classe segundo estrutura básica do PROV-DM ao qual estão classificados. Esta correlação está apoiada nas classificações descritas por Cruz et al. [2].

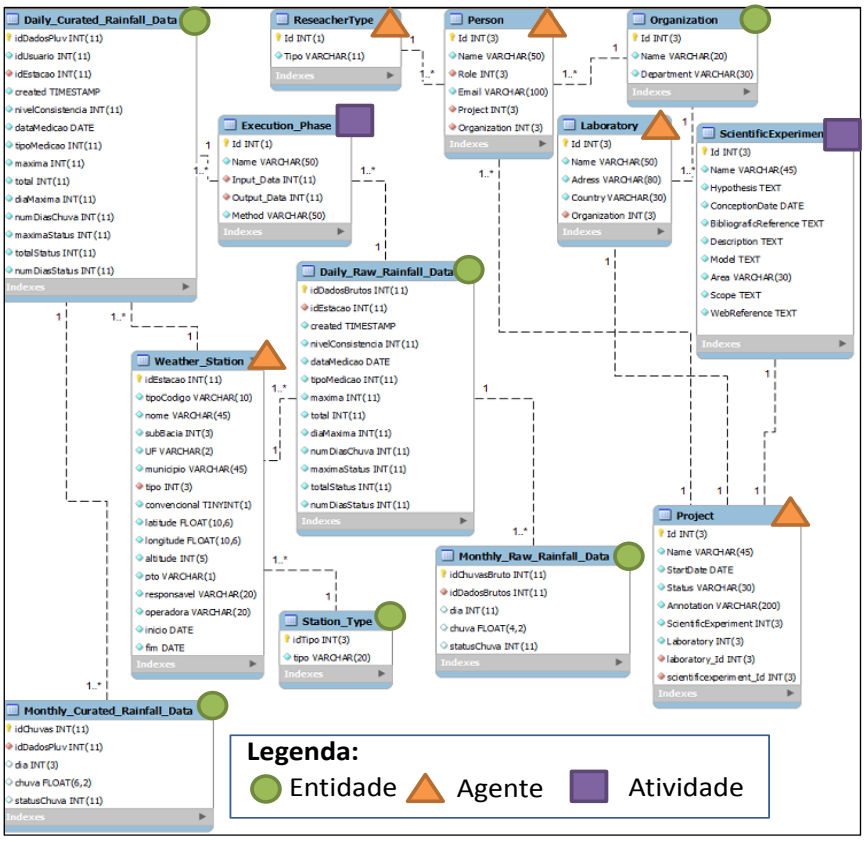

Figura 3. Representação do banco de dados.

4. Conclusão e Trabalhos Futuros

Este trabalho apresentou uma abordagem semântica para o tratamento de longas séries de dados meteorológicos que incorporou técnicas de áreas distintas (processamento de grandes volumes de dados meteorológicos, ontologias de fundamentação e de proveniência). Modelou-se uma ontologia que utiliza conceitos de proveniência, meteorologia e de gerenciamento de experimentos científicos.

Investigações e experimentações adicionais sobre a abordagem ainda se fazem necessários, em especial no que diz respeito aos estudos sobre os conceitos de eventos e relações temporais. Neste caso, estudos aprofundados do fragmento UFO-B e UFO-C são necessários. Como atividades futuras serão avaliadas a incorporação de novos

conceitos correlacionados às medições de outras variáveis físicas comuns a Meteorologia.

Referências

[1] SBMET, 2011. Introdução À Meteorologia. http://sbmet.org.br/ecomac/pages/trabalhos/introducao%20a%20meteorologia.pdf.

[2] Cruz, S. M. S.; Campos, M. L. M.; Mattoso, M., 2012. A Foundational Ontology to Support Scientific Experiments. Disponível em: ceur-ws.org/Vol-728/paper6.pdf

[3] Moreau, L.; Missier, P., 2013. PROV-DM: The PROV Data Model", W3C, www.w3.org/TR/prov-dm/.

[4] Guizzardi, G., 2005. Ontological Foundations for Structural Conceptual Models, PhD Thesis, University of Twente, Netherlands.

[5] Smolna, R.; Erwitt. J., 2012. The Human Face of Big Data. 1st Edition

[6] Papathomas, T. V.; Schiavone, J. A., Julesz, B., 1988. Applications of computer graphics to the visualization of meteorological data. XV SIGGRAPH. 327-334.

[7] Lemos Filho, G. R.; Precinoto, R. S.; Correia, T. P.; Santos, E. O.; Lyra, G. B.; Cruz, S. M. S., 2013. Assimilação, Controle de Qualidade e Análise de Dados de Meteorológicos Apoiados por Proveniência, VII e-science Workshop, XXXIII CSBC.

[8] Barbosa, T. M. S.; Cruz, S. M. S., 2013. Uma Abordagem de Gerenciamento Semântico de Experimentos Meteorológicos em Pluviometria. VII e-science Workshop, XXXIII CSBC.

[9] Freire, J.; Koop, D.; Santos, E.; Silva, C.T., 2008. Provenance for computational tasks: A survey. Computing in Science & Engineering 10 (3), 11-21.

[10] Guizzardi, G.; Halpin, T., 2008. Ontological foundations for conceptual modeling. Applied Ontology, v. 3, pp. 1-12.

[11] Moreau, L. et al., 2010. Open Provenance Model (OPM) OWL Specification.http://openprovenance.org/model/opmo.

[12] Atemezing et al., 2012. Transforming meteorological data into linked data. Semantic Web. Vol. 4. N. 3, pp 285-290. IOS Press.

[13] SSN - Semantic Sensor Network Ontology, 2005. http://purl.oclc.org/NET/ssnx/ssn.